

Recap

P-values based on multi sample splitting

avoid “double dipping” using the data twice for variable selection and using statistical inference (tests, confidence intervals) afterwards

~> sample splitting

multiple sample splitting is much more reliable and statistically better than splitting once – see later today

Fixed design linear model

$$Y = X\beta^0 + \varepsilon$$

split the sample into two parts I_1 and I_2 of equal size $\lfloor n/2 \rfloor$

- ▶ use (e.g.) Lasso to select variables based on I_1 : $\hat{S}(I_1)$
- ▶ perform low-dimensional statistical inference on I_2 based on data $(X_{I_2}^{(\hat{S}(I_1))}, Y_{I_2})$;
for example using the t -test for single coefficients β_j^0
if variable $j \in \hat{S}(I_1)$: use the p-value from t -test
if $j \notin \hat{S}(I_1)$: assign the p-value equal to 1

due to independence of I_1 and I_2 , this is a “valid” strategy
(but see later)

Validity of the (single) data splitting procedure

consider testing $H_{0,j} : \beta_j^0 = 0$ versus $H_{A,j} : \beta_j^0 \neq 0$

assume Gaussian errors for the fixed design linear model:

thus, use the t -test on the second half of the sample I_2 to get a p-value

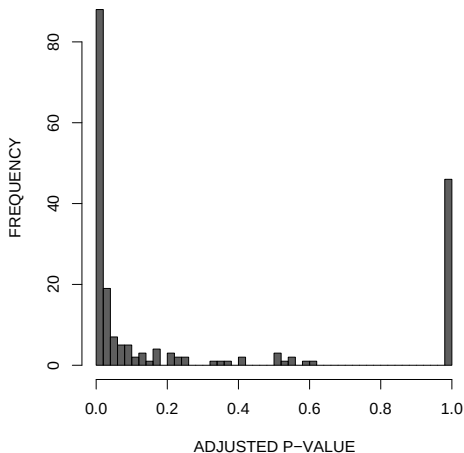
$P_{\text{raw},j}$ from t -test based on $X_{I_2}^{(\hat{S}(I_1))}$, Y_{I_2}

$P_{\text{raw},j}$ is a valid p-value (controlling type I error) for testing $H_{0,j}$
if $\hat{S}(I_1) \supseteq S_0$, i.e., the screening property holds

a p-value lottery depending on **the random split** of the data

motif regression $n = 287, p = 195$

different $P_{\text{raw},j}$ for fixed j but different splits of the data



~> should aggregate/average over multiple splits!